KCDH Interdisciplinary Dual Degree Projects - 2021-2022

Addressing Class Imbalance in Medical Image Data: Finding Needle in a Haystack

- Supervisor: Prof. Ganesh Ramakrishnan
- Co-supervisor: Prof. Rishabh Iyer

Current state-of-the-art active learning algorithms (www.arxiv.org/abs/1906.03671, www.arxiv.org/abs/2012.10630, www.cse.iitb.ac.in) do not often work well in cases where there is a class imbalance (e.g. the cancerous cases are much lower compared to the non-cancerous cases), distribution shift (e.g. training with data from one ethnicity/group and testing on another ethnicity), and out of distribution examples (e.g. unseen classes in the unlabeled set). All of these issues are often present in real-world medical imaging datasets. We will develop new active learning algorithms, which can effectively address the issues pointed out above. We will achieve this by building upon the supervisors' work on active learning using submodular mutual information functions www.arxiv.org/abs/2103.00128. To get a glimpse of the datasets and challenges in this space, look at www.grand-challenge.org. As for the coding platform, we will develop it on www.decile.org - our homegrown data efficient machine learning platform. More specifically, we will build on the DISTIL (www.github.com/decile-team/distil) - an open-source platform for Deep diversified interactive Learning with several Jupyter notebooks and video tutorials (such as www.youtube.com).

The overall goal of the project is Data and Cost-Efficient Deep Learning for Medical Image Classification. The advancement of machine learning and deep learning is creating a big impact in several domains. One such important domain is to build machine learning classifiers to effectively complement human doctors and radiologists in detecting various diseases (e.g. forms of cancer) from various medical images, e.g. X-rays, CT scans, and MRI images. Examples of applications range from cancer tumor detection, medical image segmentation, to detection of Alzheimer's and Parkinson's disease. Despite all the amazing progress of deep learning, one main challenge of these approaches is that deep learning models are extremely data-hungry and require several tens of thousands of images to work effectively. Given that detecting diseases like cancers in images requires specialized skill-sets (doctors and radiologists), the cost of annotating these datasets is very high. Furthermore, it is often hard to find sufficient samples for certain rare diseases, and medical imaging data is often heavily class-imbalanced. The goal of this project is to significantly reduce the amount of labeled data required, with minimal loss in accuracy. To achieve this, we will study the role of active learning and semi-supervised learning to reduce the amount of labeled data required. Semi-supervised approaches aim at effectively using the unlabeled data in complementing the limited labeled data in learning, while data selection and active learning seek to select the most informative labeled data to improve the model performance. Preliminary results suggest that we can reduce the amount of labeled data by factors of 5x to 20x with negligible performance degradation, depending on the dataset and the choice of the algorithm.

The advisors complement each other in this project: Ganesh Ramakrishnan's and Rishabh Iyer's expertise is in active learning and semi-supervised learning and Prof. Tamil's expertise is in deep learning for medical imaging. We will facilitate a summer internship at a Healthcare technology company/organization, as part of the collaboration, subject to approval by the faculty advisors.

Advancing causal inference methods in health economics and health services research

- Supervisor: Prof. Souvik Banerjee

The proposed study will explore use of latent factor methods to obtain causal estimates of treatment effects and its applications in healthcare settings. I have already published work on this topic and would like to extend the econometric framework.

Automated feature selection for biomarker discovery from big biological data

- Supervisor: Prof. Pramod Wangikar
- Co-supervisor: Prof. Ganesh Ramakrishnan

Proteins and metabolites, the new class of biomarkers are expected to bring a paradigm shift in the diagnosis, monitoring and treatment of human disease and will make personalized medicine a reality in near future. Moreover, the next-generation biomarkers are likely to be based on the inference drawn from multiple metabolites or protein molecules rather than single measurements such as the blood glucose level that is currently used for the diagnosis of diabetes. The present project focuses on the discovery of biomarkers from big biological data involving genomics, proteomics or metabolomics. Our primary objective will be to select the best combination of biomarkers or a minimum subset of features to predict class labels such as disease vs. healthy or one category of disease vs. another.

The challenges include:

- Small amount of labelled data
- Unbalanced data with too many features and too few samples
- The need to learn to predict classes that may have only subtle differences
- High degree of inherent biological variability (unrelated to the class label) and instrument noise.

We will use a number of feature selection methods and machine learning tools together with the concepts of multitask learning to achieve the task of biomarker discovery. A large number of public domain databases are available that will be used as test cases. See a recently published paper to understand the broad objectives of the proposed project (www.frontiersin.org). You may also see literature on multitask machine learning. with special focus on data efficient machine learning (see www.decile.org). The supervisor is active in the fields of metabolomics, systems biology and big biological data analysis. The candidate must have a strong foundation in the mathematics and programming that enable AI/ML, apart from the willingness to work in cross-disciplinary areas. We will try and facilitate a summer internship at a Healthcare technology company/organization, subject to approval by the faculty advisors.

Deep active learning for medical image classification

- Supervisor: Prof. Ganesh Ramakrishnan
- Co-supervisor: Prof. Rishabh Iyer

We will develop new active learning algorithms for the specific domain of medical imaging, which can effectively combine aspects of uncertainty (picking examples to label where the current model is most confused on) and diversity (picking representative and non-redundant examples to label) to pick the most informative samples for the specific tasks at hand. Furthermore, we will also study the sample efficiency and labeling cost reduction which can be achieved by active learning. To get a glimpse of the datasets and challenges in this space, look at www.grand-challenge.org. As for the coding platform, we will develop on www.decile.org - our homegrown data efficient machine learning platform. More specifically, we will build on the DISTIL (www.github.com/decile-team/distil) - an open source platform for Deep diversified interactive Learning with several Jupyter notebooks and video tutorials (such as www.youtube.com).

The overall goal of the project is Data and Cost Efficient Deep Learning for Medical Image Classification. The advancement of machine learning and deep learning is creating a big impact in several domains. One such important domain is to build machine learning classifiers to effectively complement human doctors and radiologists in detecting various diseases (e.g. forms of cancer) from various medical images, e.g. X-rays, CT scans, and MRI images. Examples of applications range from cancer tumor detection, medical image segmentation, to detection of Alzheimer's and Parkinson's disease. Despite all the amazing progress of deep learning, one main challenge of these approaches is that deep learning models are extremely data-hungry and require several tens of thousands of images to work effectively. Given that detecting diseases like cancers in images requires specialized skill-sets (doctors and radiologists), the cost of annotating these datasets is very high. Furthermore, it is often hard to find sufficient samples for certain rare diseases, and medical imaging data is often heavily class-imbalanced. The goal of this project is to significantly reduce the amount of labeled data required, with minimal loss in accuracy. To achieve this, we will study the role of active learning and semi-supervised learning to reduce the amount of labeled data required. Semi-supervised approaches aim at effectively using the unlabeled data in complementing the limited labeled data in learning, while data selection and active learning seek to select the most informative labeled data to improve the model performance. Preliminary results suggest that we can reduce the amount of labeled data by factors of 5x to 20x with negligible performance degradation, depending on the dataset and the choice of the algorithm.

The advisors complement each other in this project: Ganesh Ramakrishnan's and Rishabh Iyer's expertise is in active learning and semi-supervised learning and Prof. Tamil's expertise is in deep learning for medical imaging. We will facilitate a summer internship at a Healthcare technology company/organization, as part of the collaboration, subject to approval by the faculty advisors.

Disparities in access to and utilization of healthcare services in India

- Supervisor: Prof. Souvik Banerjee

This study will examine disparities in access to and utilization of healthcare services in terms of geography (rural vs. urban), religion, and socioeconomic status using a nationally-representative dataset in India.

Implementation of an alert system for early diagnosis of Acute Kidney Injury

- Supervisor: Prof. Siuli Mukhopadhyay
- Co-supervisor: Dr. Barnali Das

Acute kidney injury (AKI) is a commonly occurring condition in patients admitted to the hospital associated with a high rate of complications, morbidity and mortality. This disease is hard to diagnose owing to its asymptomatic pathogenesis and events of AKI are often missed by physicians. There is also a widespread unawareness regarding the dangers of this disease not only in patients but also in physicians who are not nephrologists.

We intend to implement the first-ever alert system, for early diagnosis of AKI, in India. In the pre-implementation phase of the study, we will create an algorithm that will track changes in serum creatinine in all eligible patients, as per the Kidney Disease Improving Global Outcomes (KDIGO) criteria. In the post-implementation phase, we will evaluate if there are improvements in patient-related outcomes as well as clinician behavior in comparison to the pre-implementation phase findings. This study will help us identify the epidemiological trends of AKI patients in an Indian hospital setting, evaluate the potential of an alert system in diagnosing AKI early leading to early intervention and possibly mitigating the poor outcomes associated with this silent killer. This study will also set a benchmark for public health advisories in an Indian healthcare setting for preventing the adversities of an easily preventable disease.

Semi-supervised learning in Medical image data

- Supervisor: Prof. Ganesh Ramakrishnan
- Co-supervisor: Prof. Rishabh Iyer

A large amount of unlabeled data that are available in the medical imaging arena can be put to use through semi-supervision learning (SSL), where the unlabeled data is effectively used in the training procedure. Through the use of the unlabeled data, semi-supervised learning has shown accuracy improvements of around 3 - 5%, compared to supervised learning with the small labeled set (www.arxiv.org/abs/2001.07685, www.arxiv.org/abs/2008.09887). We can additionally use active learning to select the most informative samples thereby reducing the amount of labeled data required to achieve the same accuracies in SSL. To get a glimpse of the datasets and challenges in this space, look at www.grand-challenge.org. As for the coding platform, we will develop onwww.decile.org - our homegrown data efficient machine learning platform. More specifically, we might build on the SPEAR (www.github.com/decile-team/spear) - an open-source platform for Semi-supervised data programming with several Jupiter notebooks and video tutorials.

The overall goal of the project is Data and Cost-Efficient Deep Learning for Medical Image Classification. The advancement of machine learning and deep learning is creating a big impact in several domains. One such important domain is to build machine learning classifiers to effectively complement human doctors and radiologists in detecting various diseases (e.g. forms of cancer) from various medical images, e.g. X-rays, CT scans, and MRI images. Examples of applications range from cancer tumor detection, medical image segmentation, to detection of Alzheimer's and Parkinson's disease. Despite all the amazing progress of deep learning, one main challenge of these approaches is that deep learning models are extremely data-hungry and require several tens of thousands of images to work effectively. Given that detecting diseases like cancers in images requires specialized skill-sets (doctors and radiologists), the cost of annotating these datasets is very high. Furthermore, it is often hard to find sufficient samples for certain rare diseases, and medical imaging data is often heavily class-imbalanced. The goal of

this project is to significantly reduce the amount of unlabeled data required, with minimal loss in accuracy. To achieve this, we will study the role of active learning and semi-supervised learning to reduce the amount of labelled data required. Semi-supervised approaches aim at effectively using the unlabelled data in complementing the limited labelled data in learning, while data selection and active learning seek to select the most informative labelled data to improve the model performance. Preliminary results suggest that we can reduce the amount of labelled data by factors of 5x to 20x with negligible performance degradation, depending on the dataset and the choice of the algorithm.

The advisors complement each other in this project: Ganesh Ramakrishnan's and Rishabh Iyer's expertise is in active learning and semi-supervised learning and Prof. Tamil's expertise is in deep learning for medical imaging. We will facilitate a summer internship at a Healthcare technology company/organization, as part of the collaboration, subject to approval by the faculty advisors.
